PATENT
Atty. Docket No. NUAN-00800

ADAPTATION OF A SPEECH RECOGNITION SYSTEM ACROSS MULTIPLE REMOTE SESSIONS WITH A SPEAKER

Field of the Invention

The present invention relates to the field of speech recognition. More particularly, the present invention relates to the field of adaptation of a speech recognition system across multiple remote sessions with a speaker.

Background of the Invention

Speech recognition systems are known which permit a user to interface with a computer system using spoken language. The speech recognition system receives spoken input from the user, interprets the input, and then translates the input into a form that the computer system understands.

Speech recognition systems typically recognize spoken words or utterances based upon an acoustic model of a person who is speaking (the speaker). Acoustic models are typically generated based upon samples of speech. When the acoustic model is constructed based upon samples of speech obtained from a number of persons rather than a specific speaker, this is called speaker-independent modeling. When a speaker-independent model is then modified for recognizing speech of a particular person based upon samples of that person's speech, this is called adaptive modeling. When a model is constructed based solely on the speech of a particular person, this is termed speaker-dependent modeling.

Speaker-independent modeling generally enables a number of speakers to interface with the same recognition system without having obtained prior samples of the speech of the particular speakers. In comparison to speaker-independent modeling, adaptive modeling and speaker-dependent modeling generally enable a speech recognition system to more accurately recognize a speaker's speech, especially if the speaker has a strong accent, has a phone line which produces unusual channel characteristics or for some other reason is not well modeled

2

by speaker independent models.

Fig. 1 illustrates a plurality of speaker-dependent acoustic models M_1, M_2, and M_n in accordance with the prior art. For each speaker, 1 through n, a corresponding speaker-dependent acoustic model M_1 through M_n, is stored. Thus, speech 10 of speaker 1 is recognized using the model M_1 and the results 12 are outputted. Similarly, speech 14 of speaker 2 is recognized using the model M_2 and the results 16 are outputted. And, speech 18 of speaker n is recognized using the model M_n and the results are outputted.

A speech recognition application program called NaturallySpeaking™, which adapts to a particular user, is available from Dragon Systems, Inc. This application program enables a user to enter text into a written document by speaking the words to be entered into a microphone attached to the user's computer system. The spoken words are interpreted and translated into typographical characters which then appear in the written document displayed on the user's computer screen. To adapt the application program to the particular user and to background noises of his or her environment, the user is asked to complete two initial training sessions during which the user is prompted to read textual passages aloud. A first training session requires that the user read several paragraphs aloud, while a second training session requires 25 to 30 to minutes for speaking and 15 to 20 minutes for processing the speech.

Other speech recognition systems are known which adapt to an individual speaker based upon samples of speech obtained while the speaker is using the system, without requiring a training session. The effectiveness of this type of adaptation, however, is diminished when only a small sample of speech is available.

Speech recognition systems are known which provide a telephonic interface between a caller and a customer service application. For example, the caller may obtain information regarding flight availability and pricing for a particular airline and may purchase tickets utilizing spoken language and without requiring assistance from an airline reservations clerk. Such customer service applications are typically intended to be accessed by a diverse population of callers and with various background noises. In such applications, it would be

impractical to ask the callers to engage in a training session prior to using the customer service application. Accordingly, an acoustic model utilized for such customer service applications must be generalized so as to account for variability in the speakers. Thus, speaker-independent modeling is utilized for customer service applications. A result of using speaker-independent modeling is that the recognition system is less accurate than may be desired. This is particularly true for speakers with strong accents and those who have a phone line which produces unusual channel characteristics.

Therefore, what is needed is a technique for improving the accuracy of speech recognition for a speech recognition system.

Summary of the Invention

The invention is a method and apparatus for adaptation of a speech recognition system across multiple remote sessions with a speaker. The speaker can remotely access a speech recognition system, such as via a telephone or other remote communication system. An acoustic model is utilized for recognizing speech utterances made by the speaker. Upon initiation of a first remote session with the speaker, the acoustic model is speaker-independent. During the first remote session, the speaker is uniquely identified and speech samples are obtained from the speaker. In the preferred embodiment, the samples are obtained without requiring the speaker to engage in a training session. The acoustic model is then modified based upon the samples thereby forming a modified model. The model can be modified during the remote session or after the session is terminated. Upon termination of the remote session, the modified model is then stored in association with an identification of the speaker. Alternately, rather than storing the modified model, statistics that can be used to modify a pre-existing acoustic model are stored in association with an identification of the speaker.

During a subsequent remote session, the speaker is identified and, then, the modified acoustic model is utilized to recognize speech utterances made by the speaker. Additional speech samples are obtained during the subsequent session and, then, utilized to further

modify the acoustic model. In this manner, an acoustic model utilized for recognizing the speech of a particular speaker is cumulatively modified according to speech samples obtained during multiple remote sessions with the speaker. As a result, the accuracy of the speech recognizing system improves for the speaker even when the speaker only engages in relatively short remote sessions.

For each speaker to remotely access the speech recognizing system, a modified acoustic model, or a set of statistics that can be used to modify the acoustic model or incoming acoustic speech, is formed and stored along with the speaker's unique identification. Accordingly, multiple different acoustic models or sets of statistics are stored, one for each speaker.

Brief Description of the Drawings

Fig. 1 illustrates a plurality of speaker-dependent acoustic models in accordance with the prior art.

Fig. 2 illustrates a speech recognizing system in conjunction with a remote communication system in accordance with the present invention.

Fig. 3 illustrates a flow diagram for adapting an acoustic model utilized for speech recognition in accordance with the present invention.

Fig. 4 illustrates a plurality of sets of transform statistics for use in conjunction with an acoustic model in accordance with the present invention.

Fig. 5 illustrates a flow diagram for adapting an acoustic model utilized for speech recognition in accordance with an alternate embodiment of the present invention.

Detailed Description of a Preferred Embodiment

Fig. 2 illustrates a speech recognizing system 100 in conjunction with a remote communication system 150 in accordance with the present invention. The remote communication system 150 can be a telephone system (e.g., a central office, a private branch

exchange or cellular telephone system). Alternately, the remote communication system 150 can be a communication network (e.g., a wireless network), a local area network (e.g., an Ethernet LAN) or a wide area network (e.g., the World Wide Web). The speech recognition system 100 includes a processing system, such as a general purpose processor 102, a system memory 104, a mass storage medium 106, and input/output devices 108, all of which are interconnected by a system bus 110. The processor 102 operates in accordance with machine readable computer software code stored in the system memory 104 and mass storage medium 106 so as to implement the present invention. The input/output devices 108 can include a display monitor, a keyboard and an interface coupled to the remote system 150 for receiving speech input therefrom. Though the speech recognizing system 100 illustrated in Fig. 2 is implemented as a general purpose computer, it will be apparent that the speech recognizing system can be implemented so as to include a special-purpose computer or dedicated hardware circuits. In which case, one or more of the hardware elements illustrated in Fig. 2 can be omitted or substituted by another.

The invention is a method and apparatus for adaptation of a speech recognizing system across multiple remote sessions with a speaker. Fig. 3 illustrates a flow diagram for adapting an acoustic model utilized for speech recognition in accordance with the present invention. The flow diagram of Fig. 3 illustrates graphically operation of the speech recognizing system 100 in accordance with the present invention. Program flow begins in a start state 200. From the state 200, program flow moves to a state 202. In the state 202, a remote session between the speaker and the voice recognition system 100 is initiated. For example, a telephone call placed by the speaker initiates the session; in which case, the speaker is a telephone caller. Alternately, the remote session is conducted via another remote communication medium. Then, program flow moves to a state 204.

In the state 204, an identification of the speaker is obtained. For example, the speaker can be prompted to speak his or her name, enter a personal identification number (pin), enter an account number, or the like. Alternately, the speaker can be automatically identified, such

as by receiving the speaker's caller ID for a telephone call. The speaker's identification can also be authenticated utilizing voice identification techniques assuming a voice sample of the speaker has previously been obtained by the speech recognition system 100. From the state 204, program flow moves to a state 206. In the state 206, a determination is made as to whether the particular speaker is a first-time speaker or if samples of the speaker's speech have been previously obtained. This is accomplished by attempting to match the speaker's identification obtained in the state 204 to a prior entry stored in the memory 104 or mass storage 106 of the speech recognizing system 100 made in response to a prior session with the same speaker. It will be apparent that the prior entries can also be stored remotely from the speech recognition system 100, such as in a centralized database which is accessible to the speech recognition system 100 via a network connection which can be provided by a local area network or the World Wide Web.

Assuming the speaker is a first time speaker, program flow moves from the state 206 to a state 208. In the state 208, a speaker-independent model is retrieved from the memory 104 or mass storage 106 to be utilized for recognizing speech made by the speaker. The speaker-independent model is a generalized acoustic model generated based upon samples of speech taken from multiple different representative persons.

The program flow then moves to a state 210. In the state 210, the speaker-independent acoustic model retrieved in the state 208 is utilized for recognizing speech made by the speaker as the speaker interacts with the speech recognition system 100 during the remote session. For example, the speaker-independent model is utilized to recognize when the speaker wishes to obtain a flight schedule, a bank account balance, and so forth. In addition, during the state 210 samples of the speaker's speech are taken. Preferably, these samples are taken without prompting the speaker to speak certain words or phrases, as in a training session. It will be apparent, however, that the speaker can be prompted to speak certain words or phrases. In which case, prompting of the speaker is preferably performed so as to minimize inconvenience to the speaker.

7

Then program flow moves to a state 212. In the state 212, the speech recognition system 100 is modified. More particularly, the speaker-independent acoustic model utilized in the state 210 to recognize the speaker's speech is modified based upon the samples of the speaker's speech taken in the state 210, thereby forming a modified acoustic model.

In the preferred embodiment, the acoustic model is modified prior to termination of the remote session so that the modified model can immediately be put to use. Alternately, the acoustic model is modified after termination of the remote session. In the preferred embodiment, the acoustic model is modified and put to use for speech recognition during the first and subsequent remote sessions. The acoustic model can also be modified between remote sessions. Thus, the states 210 and 212 can be performed repeatedly, one after the other or concurrently, during a single session. For example, assuming a predetermined amount of speech (e.g., three seconds) is received (state 210), but the remote session has not yet been terminated, then the acoustic model can be modified (state 212) while a next predetermined amount of speech is received (state 210). Once the next predetermined amount of speech is received, the acoustic model is again modified (state 212). For simplicity of illustration, however, the states 210 and 212 are shown in Fig. 3 as occurring in a simple succession. Once the session terminates, program flow moves to a state 214.

In the state 214, a representation of the modified acoustic model, such as the modified model itself or a set of statistics that can be used to modify a pre-existing acoustic model or that can be used to modify incoming acoustic speech, is stored in the memory 104 or mass storage 106 or in a centralized network database. Note that rather than modifying an acoustic model, the present invention can be utilized to modify measurements of the speech such as features vectors to achieve the principle advantages of the present invention. It will be understood that modification of phonetic features is within the scope of the present invention and that use of the term "acoustic model" herein includes phonetic features.

Thus, in the preferred embodiment, only a set of statistics which can be used to modify a pre-existing acoustic model, is stored. For example, Fig. 4 illustrates a plurality of

sets of transform statistics for use in conjunction with a pre-existing acoustic model M in accordance with the present invention. For each speaker, 1 through n, a corresponding set of statistics X_1, through X_n, is stored. For each speaker 1 through n, the modified model can be considered to include the corresponding set of statistics X_1 through X_n together with the pre-existing model M. Only one copy of pre-existing model M need be stored.

Thus, during a subsequent telephone session, speech 300 of speaker 1 is recognized using the corresponding set of transform statistics X_1 in conjunction with the pre-existing model M to recognize the speaker's speech for forming an output 302. Similarly, speech of speaker 2 is recognized using the corresponding set of statistics X_2 and the same pre-existing model M to recognize the speaker's speech for forming the output 302. And, speech of speaker n is recognized using the corresponding set of statistics X_n and the model M to recognize the speaker's speech for forming an output 302. As a result, memory is conserved in comparison with the prior technique illustrated in Fig. 1.

The modified model, or set of statistics, is stored in association with the identification of the speaker for utilization for recognizing the speaker's speech in a subsequent session with the speaker. For example, assume that in the state 206, the speech recognition system 100 looks up the speaker's identification and determines that a sample of the speaker's speech has previously been obtained. In which case, program flow moves from the state 206 to a state 216.

In the state 216, the modified model or set of statistics stored in response to the speaker's previous remote session is retrieved from the memory 104 or mass storage 106 to be utilized for recognizing speech utterances made by the speaker during the current session. From the state 216, program flow then moves to the state 210. In the state 210, the modified acoustic model or set of statistics retrieved in the state 216 is utilized for recognizing speech utterances made by the speaker as the speaker interacts with the speech recognition system during the remote session. Additional samples of the speaker's speech are taken in the state 210 and utilized to further modify the acoustic model for the speaker.

In this manner, an acoustic model utilized for recognizing the speech of a particular speaker is cumulatively modified according to speech samples obtained during multiple remote sessions with the speaker. As a result, the accuracy of the speech recognizing system improves for the speaker across multiple remote sessions even when the remote sessions are of relatively short duration.

Fig. 5 illustrates a flow diagram for adapting an acoustic model utilized for speech recognition in accordance with an alternate embodiment of the present invention. The flow diagram of Fig. 5 illustrates graphically operation of the speech recognizing system 100 in accordance with an alternate embodiment of the present invention. Portions of Fig. 5 which have a one-to-one functional correspondence with those of Fig. 3 are given the same reference numeral and are not discussed further.

The flow diagram of Fig. 5 differs from that of Fig. 3 in that from the state 210, program flow moves to a state 400. In the state 400, a determination is made relative to the incoming speech utterance. This determination preferably assigns a confidence level related to the accuracy of the speech recognition performed in the state 210. This can be accomplished by the speech recognizing system 100 assigning each speech utterance, such as a word, a phoneme, a phrase or a sentence, a certainty or score, where the assigned certainty or score is related to the probability that the corresponding identified speech correctly corresponds to the spoken input, and, then, comparing the certainty or score to one or more predetermined thresholds. If the speech recognition confidence is consistently extremely high, there may be no need to modify (or further modify) the acoustic model for the particular speaker. By avoiding modification of the acoustic model, this saves processing capacity and memory of the speech recognition system 100 which can be devoted to other tasks. Conversely, if the speech recognition accuracy is extremely low, any modifications made to acoustic model based upon incorrectly recognized speech utterances or words is not expected to improve the accuracy of speech recognition based upon such a modified acoustic model. Accordingly, if the determination made in the state 400 suggests a high accuracy (e.g., the certainty exceeds a

first threshold) or a low accuracy (e.g., the certainty is below a second threshold that is lower than the first threshold), then program flow returns to the state 202 upon termination of the remote session. In which case, no modifications to the acoustic model are performed.

Alternately, assuming the speech recognition accuracy is determined to be moderate (e.g, the certainty falls between the first and second thresholds), then it is expected that modifications to the acoustic model will improve accuracy. In which case, program flow moves from the state 400, to the state 212. As discussed relative to Fig. 3, in the state 212, the speaker-independent acoustic model utilized in the state 210 to recognize the speaker's speech is modified based upon the samples of the speaker's speech taken in the state 210, thereby forming a modified acoustic model.

In addition, because each portion of an utterance, such as a word or a phoneme, can be associated with a different confidence level, a single utterance can have several confidence levels associated with it. Thus, if some levels are above a threshold and others are below, only those portions having a confidence level above the threshold can be used to update the model.

Note that criteria other than, or in addition to, confidence levels can be utilized for making the determination in the state 400 of whether or not to modify the acoustic model. For example, a level of available resources in the speech recognition system 100, such as a low level of available memory or available processing power, may indicate that modification of the model is undesirable.

In the state 214, a representation of the modified acoustic model, such as the modified model itself or a set of statistics that can be used to modify a pre-existing acoustic model, is stored in the memory 104 or mass storage 106 or in a centralized network database in association with the identification of the speaker for utilization for recognizing the speaker's speech in a subsequent remote session with the speaker.

In an alternate embodiment, the determination made in the state 400 can be supervised. For example, the speech recognition system 100 can inform the speaker of the word or words

it has recognized and, then, ask the speaker to verify whether the speaker's speech has been correctly recognized. Assuming the speaker confirms that the speaker's speech has been correctly recognized, then program flow moves from the state 400 to the state 212. Accordingly, the correctly identified speech utterances or words are utilized to modify the acoustic model. Conversely, if the speaker indicates that the speech utterances or words were incorrectly identified, then the acoustic model is not modified based upon such incorrectly identified speech utterances or words.

As described in relation to Figs. 2-5, an acoustic model utilized for recognizing the speech of a particular speaker is cumulatively modified according to speech samples obtained during multiple remote sessions with the speaker. As a result, the accuracy of the speech recognizing system improves for the speaker across multiple remote sessions even when the sessions are of relatively short duration.

A feature of the present invention provides an acoustic model that uniquely corresponds to each of a plurality of speakers. During a first remote session with each of the speakers, the speaker-independent acoustic model is initially utilized. This model is then modified according to speech samples taken for each particular speaker. Preferably, the model is modified during the first and subsequent remote sessions and between sessions. Each modified model is then stored in association with the corresponding speaker's identification. For subsequent remote sessions, the speech recognizing system 100 retrieves an appropriate acoustic model from the memory 104 or mass storage 106 based upon the speaker's identification. Accordingly, each acoustic model is modified based upon samples of the corresponding speaker's speech across multiple remote sessions with the speaker.

To conserve memory, acoustic models that are specific to a particular speaker can be deleted from the memory 104 or mass storage 106 when no longer needed. For example, when a particular speaker has not engaged in a remote session with the service application for a predetermined period of time, then the acoustic model corresponding to that speaker is deleted. Should the speaker initiate a remote session after deletion of the acoustic model

12

corresponding to that speaker, the speaker-independent model is initially utilized and then modified according to newly acquired samples of the speaker's speech, as described above.

According to yet another embodiment of the present invention, rather than modifying an acoustic model across a plurality of remote sessions based upon speech of an individual speaker such that the model is speaker specific, the acoustic model can be modified based upon speech of a group of speakers such that the model is speaker-cluster specific. For example, speakers from different locales, each locale being associated with a corresponding accent (or lack thereof), can be clustered and a model or set of statistics can be stored corresponding to each cluster. Thus, speakers from Minnesota can be included in a cluster, while speakers from Georgia can be included in another cluster. As an example, when the remote connection is via telephone, the speaker's telephone area code can be used to place the speaker into an appropriate cluster. It will be apparent that clusters can be based upon criteria other than locale.

The present invention has been described in terms of specific embodiments incorporating details to facilitate the understanding of principles of construction and operation of the invention. Such reference herein to specific embodiments and details thereof is not intended to limit the scope of the claims appended hereto. It will be apparent to those skilled in the art that modifications may be made in the embodiment chosen for illustration without departing from the spirit and scope of the invention.